Análisis estadístico megavariado de datos de alta densidad en diagnósticos médicos sobre riesgos de cáncer hepático

Jesús Alberto Martínez Castro

Centro de Investigación en Computación

Susana Reyes Cadena

Instituto Nacional de Rehabilitación

1. Resumen

Se describen los métodos computacionales empleados en el análisis de Componentes principales y Mínimos cuadrados parciales en estudios sobre el cáncer. Parte fundamental del presente trabajo consistió en instalar y en adecuar las clases que se ofrecen en el sistema de análisis estadístico *R* y comparar su funcionalidad con el sistema *ROOT* y programas comerciales como Matlab, evaluar su funcionalidad ante análisis demandantes (gran cantidad de datos) y comparar su nivel de ejecución en clusters de alto rendimiento instalados específicamente para ofrecer en un futuro cercano el servicio de análisis de datos Bioinformáticos a la comunidad Politécnica por parte del Centro de Investigación en Computación. Por acuerdos de confidencialidad y por criterios de comparación de resultados con otros desarrollos de software, se emplearán para este reporte exclusivamente datos accesibles al público en general, cuya fuente se describe en el texto.

2. Introducción

En décadas pasadas, ha habido un gran aumento en la cantidad de datos provenientes de secuenciaciones de ADN, esto se debe, en gran medida, al éxito que han tenido los diferentes experimentos en secuención genómica. A la fecha existen tres bases de datos (entre otras) que ofrecen aceso libre a sus bases de información, a saber:

- **EMBL** (European Molecular Biology Laboratory)
  Base mantenida por European Bioinformatics Institute in Cambridge, UK. [1].

- **GenBank**
  Mantenida por National Center for Biotechnology Information in Maryland, USA [2].

- **DDBJ** (DNA Data Bank of Japan)
  Mantenida por National Institute of Genetics in Mishima, Japan. [3].

Información obtenida de GenBank sugiere un crecimiento exponencial (ver Fig(1)) en la cantidad de datos provenientes de secuenciaciones, por esto, se necesita de nuevos métodos computacionales además de mejores algoritmos y sistemas de clasificación de datos más eficientes (ontologías). Para hacer frente a este tipo de problemas, surge la Bioinformática, la cual es el desarrollo de métodos computacionales para estudiar la estructura, función y evolución de genes proteínas y el genoma como un todo.

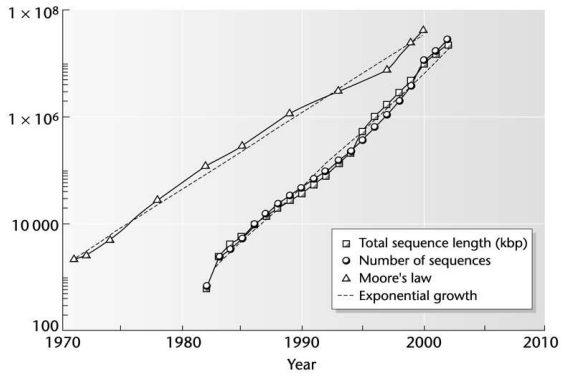

Figura 1: Velocidad de crecimiento de la cantidad de datos registrados en el GenBank sequence comparados con el número de transistores usados en computadoras personales. Fuente www.ncbi.nih.gov/Genbank/genbankstats.html

Las tendencias modernas de análisis de datos, tienen como filosofía el estudiar grupos de genes en conjunto en lugar de considerar casos individuales. Para esto, entre otras técnicas se emplea el Análisis de componentes principales (PCA por sus siglas en inglés). El PCA ofrece una manera de visualizar las características realmente importantes en un conjunto multidimensional de datos. El método construye un conjunto de coordenadas que se expresan como combinación lineal de las variables originales, de forma tal, que la máxima variabilidad entre los datos se explica en términos de pocas coordenadas.

Una técnica más para estudiar este tipo de datos, es el análisis de clusters, el cual se encarga de buscar patrones en conjuntos complejos de datos. Los análisis de clusters se subdividen en jerárquicos y directos, sin embargo, dada su construcción matemática, el análisis de clusters puede aceptar diversas interpretaciones por lo que es necesario de discusiones extensas para concluir el análisis.

La meta de este proyecto es realizar un sistema computacional que realice PCA, PLS, y análisis de clusters. Se probará el sistema con datos provenientes de estudios para deteccción de cancer. Se enfatiza que Jesús Martínez no tomó dato alguno y que solo es responsable por el diseño e implementación del software de análisis, el cual se describe en la siguiente sección.

3. Métodos y materiales

Se empezó el proyecto evaluando las versiones comerciales de software que realizan este tipo de análisis, poniéndose especial atención a *Matlab* con su *Toolkit* de Bioinformática.

Los resultados de ejecución de *Matlab* se compararon con desarrollos propios en C++ en gcc, gcc-3.4, gcc-4.2, dadas las diferencias en el manejo de flotantes y dado que gcc-3.4 emplea rutinas en F77 a diferencia de gcc-3.4 donde se decidió emplear F95.

El método de evaluación consistió en comparar los ROOTMARKS en corridas y criterios de primera y segunda derivada (ver manual de GLOBUS TOOLKIT).

En las pruebas realizadas, *Matlab* obtuvo 2237 ROOTMARKS, nuestros desarrollos obtuvieron 2149 ROOTMARKS, por lo que concluimos que las rutinas numéricas ofrecidas por gcc-3.4 eran las mejores para el trabajo de análisis.

Por comparación, un procesador Pentium IV a 2.4 GHz se evalua con 600 ROOTMARKS.

A continuación, se procedió a descargar los datos (en diferentes formatos de las bases gratuitas. Se decidió en su momento usar los datos en formato simple con separaciones de tabulador, esto para no depender de ninguna compañia privada que al inicio ofrezca un sistema gratuito y que de pronto decida cobrar por el servicio.

Para nuestro caso, los datos podrían ser visualizados crudos en cualquier hoja de cálculo o procesador de textos.

Una vez descargados los datos, se procedió a escalarlos y a reescribirlos en términos del nuevo origen como se ilustra en la Fig(2).

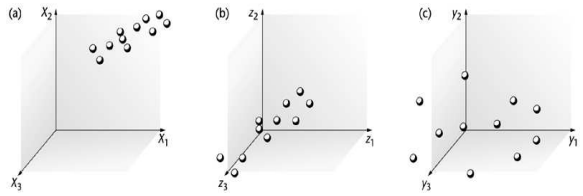

Figura 2: Esquematización de expresión de datos en términos de ejes principales. (a)Datos originales. (b)Datos escalados y centrados en el origen. (c)Rotación sobre ejes principales.

Se consideró que N el número de objetos en nuestra muestra y P el número de columnas (sujetos de estudio, muestras, propiedades). Se procedió a calcular la matrix de coeficientes de correlación

$$C_{jk} = \frac{1}{N\sigma_j\sigma_k}\sum_i (X_{ij} - \mu_j)(X_{jk} - \mu_k) = \frac{1}{N}\sum_j z_{ij}z_{jk} \tag{1}$$

Los coeficientes estan entre -1 y 1 y se comprobó que los vectores V_j cumplen con

$$\sum_j v_{nj}C_{jk} = \lambda_n v_{nk} \tag{2}$$

De nuestro análisis se tomó que el primer componente principal es aquel con el eigenvalor mayor, los siguientes se tomaron en orden decreciente.

$$\lambda_1 = 3{,}43 \qquad\qquad \lambda_2 = 1{,}61 \tag{3}$$

Se calculó también la varianza a lo largo del n^{th} eje, la cual por construcción se puede demostrar que es igual al eigenvalor correspondiente

$$\frac{1}{N}\sum_i y_{in}^2 = \frac{1}{N}\sum_i\sum_j\sum_k v_{nj}z_{ij}v_{nk}z_{ik} = \sum_j\sum_k v_{nj}C_{jk}v_{nk} = \sum_k \lambda_n v_{nk}^2 = \lambda_n \tag{4}$$

La fracción de la varianza total representada por los primeros dos componentes es

$$\frac{\lambda_1 + \lambda_2}{P} = \frac{3{,}43 + 1{,}61}{8} = {,}63 \tag{5}$$

Se desarrolló también un módulo que ofrece la posibilidad de analisis de clusters con las siguientes restricciones:

Distancia entre vectores

$$|z_i| = (\sum_k z_{ik}^2)^{\frac{1}{2}} \tag{6}$$

donde el ángulo esta dado por (ver geometría en Fig(3))

$$\cos\theta_{ij} = \frac{\sum_k z_{ik} z_{jk}}{|z_i||z_j} \tag{7}$$

Medida alternativa de similaridad considerada

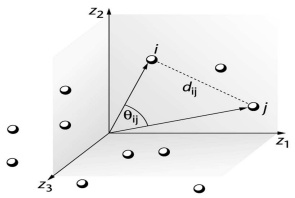

Figura 3: Representación de cada dato como vector en el espacio de ejes principales (espacio multidimensional).

$$R_{ij} = \frac{1}{Ps_i s_j} \sum_k (z_{ij} - m_i)(z_{jk} - mj) \tag{8}$$

distancia considerada

$$d_{ij} = (\sum_k (z_{ik} - z_{jk})^2)^{\frac{1}{2}} \tag{9}$$

Función alternativa de cluster no jerarquizado

$$I_2 = \sum_A \sqrt{\sum_{i,j \in A} S_{ij}} \tag{10}$$

$$d_{iA}^2 = \sum_j (z_{ij} - \mu_{Aj})^2 \tag{11}$$

Función de error

$$E = \sum_A \sum_{i \in A} d_{iA}^2 \tag{12}$$

Las rutinas desarrolladas se compararon con las que ofrece el sistema R [1]. Encontrándose mejor ejecución en R pero sujeto a los parches que se ofrecen en el sitio web.

El sistema final puede ejecutarse como script tanto en *ROOT* [2] como en R de manera independiente. Se sugiere trabajar en *ROOT* e instalar las rutinas de conexión con R dado que *ROOT* ya tiene conexiones semiautomáticas a Clusters de cómputadoras [4], en caso de que la cantidad de datos solicite ejecuciones más demandantes.

El programa en su módulo de PCA se comparó con uno ofrecido libremente [3] obteniéndose 942 ROOTMARKS para el desarrollo propio y 817 ROOTMARKS para el ofrecido en la web. Se aclara que el ofrecido en la web esta escrito principalmente en JAVA y C++.

Como manejador de bases de datos se empleo mysql-5.0.27 con mysql-connector-odbc-3.51.12 ambos suministrados en el disco de instalación.

Finalmente y con la experiencia adiquirida, se propuso un modelo de datos basado en el estandar de el GenOntology Project [4].

4. Resultados

Se diseñó, desarrolló e implementó el módulo de **Software para análisis PCA y PLC de datos**, el cual se anexará al software **Contador inteligente de células sanguíneas** (software desarollado con apoyo del IPN en proyecto anterior). Ambos softwares estarán disponibles de manera libre y gratuita en la página web del investigador, a la cual se podrá acesar por medio de la pagina del CIC

(http://www.cic.ipn.mx)

El estudio con el software permitirá por un lado, entender un poco más la evolución de la enfermedades analizadas por microarreglos (ver descripción de datos de entrada en el manual de instalación de módulos extras) y por otro permitir la clasificación ontológica de los datos.

[1] http://www.r-project.org/
[2] http://root.cern.ch/
[3] http://astro.u-strasbg.fr/ fmurtagh/mda-sw/
[4] http://www.geneontology.org/

Los datos y resultados generados se clasificarán e indexarán gracias a su estructura ontológica (ver manual del usuario) en una base de conocimiento que podrá ser aplicada a sistemas de control de diversos tipos de cáncer .

Como parte de un futuro proyecto de investigación se espera acoplar un módulo extra para poder realizar los anaálisis desde dispositivos móviles como PDA y/o celulares.

5. Impacto

El software **Contador inteligente de células sanguíneas**, desarrollado previamente en proyecto IPN, se encuentra en pleno uso en el Instituto Nacional de Rehabilitación, hospital de la Secretaría de Salud desde Agosto de 2008.

El software **Software para análisis PCA y PLC de datos** se usa en etapa de prueba de prototipo en el Instituto Nacional de Rehabilitación, desde Noviembre de 2008.

Se anexa constacia de uso de software y prueba de prototipo en los comprobantes que se entregarán a la Dirección del Centro de Investigación en Computación del IPN.

En Febrero de 2009 se propondrá extender el uso de ambos desarrollos a hospitales del Seguro Social del interior de la República, para lo cual se propondrá un plan integral que contemple capacitación y adecuación para necesidades específicas.

El hospital que se propone como piloto es el hospital de Iguala Guerrero, con quien ya se ha venido interactuando desde hace más de tres años con apoyo de proyectos IPN.

Se menciona también que ambos desarrollos fueron presentados entre las propuestas de colaboración México-Bélgica en el Groupe Interdisciplinaire de Genoproteomique Appliquee (GIGA) de la Universidad de Liege en Liege en Bélgica (se anexa constancia).

Los desarrollos impactan también en el sector educativo pues facilitan el entendimiento matemático del estudiante.

Referencias

[1] Stoesser G et al. The embl nucleotide sequence database: Major new developments. *Nucleic Acids Research.*, 31:17–22, 2003.

[2] Benson D.A et al. *Nucleic Acids Research.*, 31:23–27, 2003.

[3] Mizayaki et al. Dna databank of japan (ddbj) in xml. *Nucleic Acids Research.*, 31:13–16, 2003.

[4] Plaszczak P. *Grid Computing: The Savvy Manager's Guide (The Savvy Manager's Guides)* . Morgan-Kaufman, 2005.